MASSACHUSETTS
INSTITUTE of TECHNOLOGY
LIBRARIES

HB31
.M415
99-06
(27)

DEWEY

working paper
department
of economics

TWO-STEP ESTIMATION, OPTIMAL MOMENT CONDITIONS, AND SAMPLE SELECTION MODELS

Whitney K. Newey
James L. Powell

No. 99-06 February, 1999

massachusetts
institute of
technology

50 memorial drive
cambridge, mass. 02139

WORKING PAPER
DEPARTMENT
OF ECONOMICS

TWO-STEP ESTIMATION, OPTIMAL MOMENT CONDITIONS, AND SAMPLE SELECTION MODELS

Whitney K. Newey
James L. Powell

No. 99-06 February, 1999

MASSACHUSETTS
INSTITUTE OF
TECHNOLOGY

50 MEMORIAL DRIVE
CAMBRIDGE, MASS. 02142

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
APR 14 1999
LIBRARIES

TWO-STEP ESTIMATION, OPTIMAL MOMENT CONDITIONS, AND SAMPLE SELECTION MODELS

by

Whitney K. Newey[1]
Department of Economics
MIT

and

James L. Powell
Department of Economics
UC Berkeley

January, 1992

Revised, December 1998

Abstract: Two step estimators with a nonparametric first step are important, particularly for sample selection models where the first step is estimation of the propensity score. In this paper we consider the efficiency of such estimators. We characterize the efficient moment condition for a given first step nonparametric estimator. We also show how it is possible to approximately attain efficiency by combining many moment conditions. In addition we find that the efficient moment condition often leads to an estimator that attains the semiparametric efficiency bound. As illustrations we consider models with expectations and semiparametric minimum distance estimation.

JEL Classification: C10, C21

Keywords: Efficiency, Two-Step Estimation, Sample Selection Models, Semiparametric Estimation,

[1]Department of Economics, MIT, Cambridge, MA 02139; (617) 253-6420 (Work); (617) 253-1330 (Fax); wnewey@mit.edu (email). The NSF provided financial support for the research for this project.

Digitized by the Internet Archive
in 2011 with funding from
Boston Library Consortium Member Libraries

http://www.archive.org/details/twostepestimatio00newe

1. Introduction

Two step estimators are useful for a variety of models, including sample selection models and models that depend on expectations of economic agents. Estimators where the first step is nonparametric are particularly important, having many applications in econometrics, and providing a natural approach to estimation of parameters of interest. The purpose of this paper is to derive the form of an asymptotically efficient two step estimator, *given* a first step estimator.

The efficient estimator for a given first step nonparametric estimator will often be fully efficient, attaining the semiparametric efficiency bound for the model, as we show for some sample selection models considered by Newey and Powell (1993). Full efficiency occurs because the first step is just identified, analogous to the efficiency of a limited information estimator of a simultaneous equation when all the other equations are just identified. An analogous result for two-step parametric estimators is given in Crepon, Kramarz, and Trognon (1997), where optimal estimation in the second step leads to full efficiency if the first step is exactly identified.

We will first give some general results that characterize second step estimators that are efficient in a certain class and consider construction of estimators that are approximately efficient. We then derive the form of efficient estimators in several specific models, including conditional moment restrictions that depend on functions of conditional expectations and sample selection models where the propensity score (i.e. the selection probability) is nonparametric. We also describe how an approximately efficient estimator could be constructed by optimally combining many second step estimating equations.

Throughout the paper we rely on the results of Newey (1994) to derive the form of asymptotic variances and make efficiency comparisons. Those results allow us to sidestep regularity conditions for asymptotic normality and focus on the issue at hand, which is the form of an efficient estimator. In this approach we follow long standing econometric

practice where efficiency comparisons are made without necessarily specifying a full set of regularity conditions. Of course, we could give general regularity conditions for specific estimators (e.g. as in Newey (1994) for series estimators or Newey and McFadden (1994) for kernel estimators), but this would detract from our main purpose.

As an initial example, consider the following simple model in which the conditional mean of a dependent variable y given some conditioning variable x is proportional to its conditional standard deviation:

$$y = \beta_0\sigma(x) + u, \quad E[u|x] = 0, \quad \sigma^2(x) = Var(y|x). \tag{1}$$

Given a sample $\{(y_i, x_i')', i = 1, \ldots, n\}$ of observations on y and x, one type of estimator for β_0 would be an instrumental variables (IV) estimator replacing $\sigma(x)$ with a nonparametric estimator $\hat{\sigma}(x)$ and using an instrument $a(x)$ to solve the equation

$$\sum_{i=1}^{n} a(x_i)[y_i - \beta\hat{\sigma}(x_i)] = 0 \tag{2}$$

for $\hat{\beta} = [\sum_{i=1}^{n} a(x_i)\hat{\sigma}(x_i)]^{-1}\sum_{i=1}^{n} a(x_i)y_i$. For example, the least squares estimator would have $a(x) = \hat{\sigma}(x)$. If the data generating process and the nonparametric estimator $\hat{\sigma}(x)$ are sufficiently regular, so that $\hat{\beta}$ is root-n consistent and asymptotically normal, then the formulae given in Newey (1994) can be used to derive the following form of the asymptotic distribution for $\hat{\beta}$:

$$\sqrt{n}(\hat{\beta} - \beta_o) \xrightarrow{d} N(0, \{E[a(x)\sigma(x)]\}^{-2} \cdot E[\zeta^2 a(x)^2]), \tag{3}$$

$$\zeta = u - [2\sigma(x)]^{-1}\{u^2 - \sigma(x)^2\}.$$

The asymptotic variance of this estimator is that of an IV estimator with instrument $a(x)$, regressor $\sigma(x)$ and residual ζ. The efficient choice of instrument, as in Chamberlain (1987) is $\sigma(x)/\omega(x)$ for $\omega(x) = E[\zeta^2|x]$. The novel feature of this optimal

instrument is that it depends on the inverse of the conditional variance of ζ rather than the original residual u. If $v = u/\sigma(x)$ is independent of x then $\zeta = \sigma(x)[v - (v^2-1)/2]$, so that $E[\zeta^2|x]$ is proportional to $\sigma^2(x)$, and the the best instrument is $1/\sigma(x)$, the same as if the first stage estimation were not accounted for. In general though, it is necessary to account for the first-stage estimator in forming the optimal instrument for the second stage.

The best IV estimator is weighted least squares with weight $\omega(x)^{-1}$. As in Newey (1993), estimation of the optimal instrument should not affect the asymptotic variance, so the weighted least squares estimator, so for $\hat{\omega}(x)$ that is suitably well behaved,

$$\tilde{\beta} = [\sum_{i=1}^{n} \hat{\omega}(x_i)^{-1}\hat{\sigma}(x_i)^2]^{-1}\sum_{i=1}^{n} \hat{\omega}(x_i)^{-1}\hat{\sigma}(x_i)y_i \qquad (4)$$

should be efficient. Alternatively, as we discuss below, an approximately efficient estimator could be constructed by GMM estimation with moment conditions $A(x)[y-\beta\hat{\sigma}(x_i)]$, where $A(x)$ is some vector of approximating functions.

2. General Methods

To describe the general class of estimators we consider, let $m(z,\theta,\alpha)$ denote a vector of functions, where z denotes a single data observation, θ is a parameter vector, and α is an unknown function. Suppose that the moment restrictions

$$E[m(z,\theta_0,\alpha_0)] = 0 \qquad (5)$$

are satisfied, where the "0" subscript denotes true values. This moment restriction and an estimator $\hat{\alpha}$ of α_0 can be used to construct an estimator $\hat{\theta}$ of θ_0, by solving the equation

$$\sum_{i=1}^{n} m(z_i,\theta,\hat{\alpha})/n = 0. \tag{6}$$

The class of estimators we consider are of this form, where m is restricted to be in some set of feasible moment functions, and α may depend on m. In the example of Section 1, $z = (y, x')'$, $\theta = \beta$, $\alpha = \sigma$ denotes the conditional standard deviation function, and $m(z,\theta,\alpha) = \sigma(x)[y - \beta\sigma(x)]$.

To characterize the optimal $m(z,\beta,\alpha)$ in some class, i.e. the one where $\hat{\theta}$ has the smallest asymptotic variance, we need the asymptotic distribution of $\hat{\theta}$. In general, the asymptotic variance will depend on the form of $\hat{\alpha}$, but for the moment, we will simply assume that for each $m(z,\theta,\alpha)$ there is an associated function $u_m(z)$ such that

$$\sum_{i=1}^{n} m(z_i,\theta_0,\hat{\alpha})/\sqrt{n} = \sum_{i=1}^{n} u_m(z_i)/\sqrt{n} + o_p(1). \tag{7}$$

Here $u_m(z)$ is the influence function of $\sum_{i=1}^{n} m(z_i,\theta_0,\hat{\alpha})/n$ with the term $u_m(z)-m(z,\theta_0,\alpha_0)$ accounting for the presence of $\hat{\alpha}$. When α is nonparametric the results of Newey (1994) can be used to derive $u_m(z)$. If equation (7) holds along with other regularity conditions then for $H_m = \partial E[m(z,\theta,\alpha_0)]/\partial\theta|_{\theta=\theta_0}$,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0,\ H_m^{-1}E[u_m(z)u_m(z)']H_m^{-1\prime}). \tag{8}$$

The efficient two-step estimator we will consider is one where m minimizes the asymptotic variance $H_m^{-1}E[u_m(z)u_m(z)']H_m^{-1\prime}$.

A sufficient condition for $\bar{m}(z,\theta,\alpha)$ to minimize the asymptotic variance of $\hat{\theta}$ is that for all m,

$$H_m = E[u_m(z)u_{\bar{m}}(z)']. \tag{9}$$

When this equation holds then by Newey and McFadden (1994), a lower bound on the asymptotic variance will be $(E[u_{\bar{m}}(z)u_{\bar{m}}(z)'])^{-1}$, and will be attained when $m = \bar{m}$. Equation (9) is analogous to the generalized information matrix equality in parametric

models, and similar equations have been used by Hansen (1985a), Hansen, Heaton, and Ogaki (1988), and Bates and White (1993) to find efficient estimators. Here we use this equation to derive the optimal choice of a second step estimator.

This characterization of an efficient two-step estimator can be used to derive the optimal estimator in the initial example. In that example, $u_m(z) = a(x)\zeta$ and $H_m = -E[a(x)\sigma(x)]$. Also, the choice of m reduces to a choice of instrument a and equation (9) is $E[a(x)\sigma(x)] = E[a(x)\zeta^2\bar{a}(x)] = E[a(x)\omega(x)\bar{a}(x)]$. A solution to this equation, that is hence an optimal instrument is $\bar{a}(x) = \omega(x)^{-1}\sigma(x)$.

Construction of an efficient estimator can often be based on the solution $\bar{m}(z,\beta,\alpha)$ to equation (9). Although $\bar{m}$ may depend on unknown functions other then α, they can often be replaced by parametric or nonparametric estimators without affecting the efficiency of $\hat{\theta}$, e.g. as in Newey (1993). Estimators which are efficient for some restricted class of unknown distributions, referred to as *locally efficient* here, can be constructed by using finite dimensional parameterizations of unknown components of the optimal moment function. Estimators which are efficient for all distributions can be constructed by using nonparametric methods to estimate unknown components. In the examples to follow we will discuss various estimators of the optimal moment functions, which will result in efficiency under appropriate regularity conditions.

A general approach to efficient estimation, which is useful when $\bar{m}$ is complicated and it is hard to form an explicit estimate, is to use the efficient generalized method of moments estimator based on "many" moment functions. This approach has been considered by Beran (1976), Hayashi and Sims (1983), Chamberlain (1987), and Newey (1993). Under a "spanning" condition, this approach will result in an estimator that is approximately efficient, in the sense that as the number of moments grows, the asymptotic variance of the estimator approaches that of the optimal estimator.

To be precise, consider a $J \times 1$ vector of functions $m_J(z,\theta,\alpha)$ (where α may depend on J). Suppose that for some $J \times 1$ vector $u_J(z)$, equation (7) is satisfied with m_J and u_J replacing m and u_m respectively, and let $\hat{V}_J$ denote an estimator

of $V_J = E[u_J(z)u_J(z)']$ (e.g. $\hat{V}_J = \sum_{i=1}^n \hat{u}_J(z_i)\hat{u}_J(z_i)'/n$ for an estimator $\hat{u}_J(z)$). An optimal GMM estimator based on the moment vector m_J is

$$\hat{\theta}_J = \text{argmin}_{\theta\in\Theta} \sum_{i=1}^n m_J(z_i,\theta,\hat{\alpha})' \hat{V}_J^{-1} \sum_{i=1}^n m_J(z_i,\theta,\hat{\alpha}). \tag{10}$$

An alternative one-step version is, for $\hat{H}_J = \sum_{i=1}^n \partial m_J(z_i,\hat{\theta},\hat{\alpha})/\partial\theta/n$,

$$\tilde{\theta}_J = \hat{\theta} - (\hat{H}_J'\hat{V}_J^{-1}\hat{H}_J)^{-1}\hat{H}_J'\hat{V}_J^{-1}\sum_{i=1}^n m_J(z_i,\hat{\theta},\hat{\alpha}) \tag{11}$$

where $\hat{\theta}$ is an initial estimator. As usual, the one-step estimator is asymptotically equivalent to its full optimization counterpart. Both estimators will have asymptotic variance $(H_J'V_J^{-1}H_J)^{-1}$, which can be estimated by $(\hat{H}_J'\hat{V}_J^{-1}\hat{H}_J)^{-1}$.

As J gets larger the asymptotic variance of this estimator will approach the lower bound, if linear combinations of $u_J(z)$ can approximate the optimal influence function $u_{\overline{m}}(z)$ in mean square, as shown by the following result.

Theorem 2.1: Suppose that there is $\overline{m}$ such that $E[u_{\overline{m}}(z)u_{\overline{m}}(z)']$ is nonsingular, $H_m = E[u_m(z)u_{\overline{m}}(z)']$ for all feasible m, and there are conformable constant matrices C_J, $J = 1, 2, \ldots$ such that $E[\|u_{\overline{m}}(z)-C_J u_J(z)\|^2] \longrightarrow 0$ as $J \longrightarrow \infty$. Then $(H_J'V_J^{-1}H_J)^{-1} \longrightarrow (E[u_{\overline{m}}(z)u_{\overline{m}}(z)'])^{-1}$ as $J \longrightarrow \infty$.

The mean-square approximation hypothesis of this result is the spanning condition referred to above. This result falls short of an efficient estimation result, because it does not specify a way, independent of the true data generating process, such that J grow with the sample size so that $\hat{\theta}_{J(n)}$ has asymptotic variance $(E[u_{\overline{m}}(z)u_{\overline{m}}(z)'])^{-1}$. It is possible to give such rates in particular problems, as in Newey (1993), but to avoid technical detail rates are not derived here. Instead, we focus on how this result suggests efficient estimation might approximately be achieved, by choosing moment functions with corresponding $u_J(z)$ that approximate $u_{\overline{m}}(z)$ in mean-square.

3. Conditional Moment Restrictions and Nonparametric Generated Regressors.

The first specific model we consider is a semiparametric instrumental variables estimator that depends on a nonparametric regression. The introductory example is a special case of this model. Also, this case includes estimators that have been considered by Ahn and Manski (1993) and Rilstone (1989). Let $a(x)$ denote a vector of instrumental variables, $\rho(z,\theta,\alpha)$ a residual that depends on a function α, where

$$E[\rho(z,\theta_0,\alpha_0)|x] = 0, \quad \alpha_0(w) = E[d|w]. \tag{12}$$

Then the moment vectors will consist of a vector of instrumental variables $a(x)$ multiplying the residual,

$$m(z,\theta,\alpha) = a(x)\rho(z,\theta,\alpha).$$

The optimality problem here is finding the set of instrumental variables that minimizes the asymptotic variance.

To derive the optimal instruments we need to account for the nonparametric estimation, which can be done by imposing the following condition. Let $\alpha(w,\gamma)$ denote a parametric specification for the conditional expectation, satisfying regularity conditions, along with the conditional moment vector, so that the derivatives in the following condition exist.

Assumption 3.1: There is $\delta(w)$ such that for all $\alpha(w,\gamma)$ with $\alpha_0(w) = \alpha(w,\gamma_0)$ for some γ_0, $\partial E[a(x)\rho(z,\theta_0,\alpha(\gamma))]/\partial\gamma\,|_{\gamma=\gamma_0} = E[a(x)\delta(w)\partial\alpha(w,\gamma_0)/\partial\gamma]$.

This condition leads to correction terms for estimation of α of the simple form $E[a(x)|w]\delta(w)[d - \alpha_0(w)]$, as derived in Newey (1994). The value of this approach is that it is based on a simple derivative calculation that is easy to apply. For instance,

in the initial example, where $\alpha_0(x) = E[(y^2,y)'|x]$ and $\rho(z,\theta_0,\alpha) = y - \beta\{\alpha_2(x) - [\alpha_1(x)]^2\}^{1/2}$, Assumption 3.1 is satisfied with $\delta(w) = -\beta_0[2\sigma_0(x)]^{-1}(1,-2E[y|x])$, leading to the correction term given earlier.

The first optimal instrument question we address is for the case where the instruments x are a subset of the first stage regressors. Let $D(x) = E[\partial\rho(z,\theta_0,\alpha_0)/\partial\theta|x]$, and for now assume that ρ is a scalar.

Theorem 3.1: If $x \subseteq w$ and Assumption 3.1 is satisfied, then

$$u_m(z) = a(x)\zeta, \quad \zeta = \rho(z,\theta_0,\alpha_0) + \delta(w)[d - g_0(w)],$$

and the choice of instruments that minimizes the asymptotic variance is

$$\bar{a}(x) = D(x)'(E[\zeta^2|x])^{-1}.$$

An interesting interpretation of these optimal instruments follows upon noting the form of the influence function, $u_m(z) = a(x)\zeta$, which is the instrumental variables times an "adjusted residual" ζ, that accounts for the presence of $\hat{\alpha}$. Consequently, the optimal instruments are the same as for an IV estimator without the first stage, except that the conditional variance $E[\zeta^2|x]$ has replaced $E[\rho^2|x]$. The reason that the optimal instruments have this simple form is that the adjusted residual fully accounts for the generated regressors.

The initial example provides one illustration of this case. Another example is the estimator of Ahn and Manski (1993). Suppose there is a binary dependent variable $y \in \{0, 1\}$, with

$$\text{Prob}(y=1|x) = \Phi(\theta_{10}[\alpha_0(x,0)-\alpha_0(x,1)] + x_1'\theta_{20}), \quad \alpha_0(w) = E[d|w], \quad w = (x,y), \quad (13)$$

where Φ is the CDF for a standard normal. Their estimator is probit, with a nonparametric estimator $\hat{\alpha}(w)$ replacing $\alpha_0(w)$. As usual for probit, this estimator is

asymptotically equivalent to an instrumental variables estimator with

$$\rho(z,\theta,\alpha) = y - \Phi(\theta_1[\alpha(x,0)-\alpha(x,1)] + x_1'\theta_2), \tag{14}$$

$$a(x) = D(x)\Omega(x)^{-1}, \quad \Omega(x) = \Phi(v)[1 - \Phi(v)], \quad D(x) = -(\alpha_0(x,0)-\alpha_0(x,1),x_1')'\phi(v),$$

where $v = \theta_{10}[\alpha_0(x,0)-\alpha_0(x,1)] + x_1'\theta_{20}$ and $\phi(v)$ is the standard normal p.d.f.. These instruments are not optimal.

To derive the optimal instruments, note that $x \subseteq w$, so that we can apply Theorem 3.1. To do so, we note that for any functions $b(x)$ and $d(w)$, $E[b(x)\{d(x,0)-d(x,1)\}]$ $= E[b(x)d(w)\{[(1-y)/(1-\Phi(v))] - [y/\Phi(v)]\}$. Then, by $\partial\rho(z,\theta_0,\alpha(\gamma))/\partial\gamma|_{\gamma=\gamma_0} =$ $-\theta_{10}\phi(v)[\partial\alpha(x,0,\gamma)/\partial\gamma - \partial\alpha(x,1,\gamma)/\partial\gamma]$, Assumption 3.1 is satisfied for $\delta(w) =$ $-\theta_{10}\phi(v)\{[1-\Phi(v)]^{-1}(1-y) - \Phi(v)^{-1}y\}$. Then, as in Theorem 3.1, $\zeta = y-\Phi(v) -$ $\theta_{10}\phi(v)\delta(w)[d-g_0(w)]$, so for $\omega(x) = E[\{d-E[d|w]\}^2|x]$, the optimal instruments are

$$\bar{a}(x) = (E[\zeta^2|x])^{-1}D(x) \tag{15}$$

$$E[\zeta^2|x] = \Phi(v)[1-\Phi(v)] + \phi(v)^2\theta_{10}^2\Phi(v)^{-1}[1-\Phi(v)]^{-1}\omega(x).$$

As in the last example, the optimal instruments are those for weighted nonlinear least squares, where the weight is the inverse of the conditional variance of an "adjusted residual," rather than the original residual.

An optimal estimator can be constructed by using estimates of the optimal instruments. As usual, estimating instruments will not affect the asymptotic variance. The optimal instruments may be estimated by substituting $\hat{\theta}_1$, $\hat{\alpha}(x)$, and $\hat{v}$ for θ_{10}, $\alpha(x)$, and v respectively in the formula for the optimal instruments, and also substituting an estimator for $\omega(x)$. A locally efficient estimator can be constructed by estimating $\omega(x)$ as the predicted value from a regression of nonparametric squared residuals $[d-\hat{\alpha}(w)]^2$ on a few functions of x. A fully efficient estimator would

require nonparametric estimation of $\omega(x)$. Alternatively, an approximately efficient estimator could be constructed from GMM estimation using many functions of x, as considered below.

Another example of Theorem 2.1 is the semiparametric panel probit estimator of Newey (1994), where y_t, $(t = 1, 2)$, are binary variables and there is an unknown function $h(x)$ such that

$$E[y_t|x] = \Phi([x_t'\beta_0 + h(x)]/\sigma_t), \quad (t = 1, 2), \quad \sigma_2 = 1.$$

Inverting the normal CDF and differencing eliminates the unknown function $h(x)$ and gives the condition $\rho(x,\theta_0,\alpha_0) = 0$, where $\alpha_0(x) = (E[y_1|x], E[y_2|x])'$, $\theta = (\beta',\sigma_1)$, and

$$\rho(x,\theta,\alpha) = \Phi^{-1}(\alpha_2(x)) - \sigma_1\Phi^{-1}(\alpha_1(x)) - (x_2-x_1)'\beta.$$

The least squares estimator in Newey (1994) is obtained by substituting a nonparametric regression estimator $\hat{\alpha}(x)$ for $\alpha_0(x)$ and minimizing $\sum_{i=1}^n \rho(x_i,\theta,\hat{\alpha})^2$. As usual for least squares, this is asymptotically equivalent to an IV estimator with instruments $a(x) = D(x)' = \partial\rho(x,\theta_0,\alpha_0)/\partial\theta = -((x_2-x_1)',\Phi^{-1}(\alpha_{10}(x)))'$.

The optimal instruments can be derived by applying Theorem 3.1. Let $\psi(a) = \phi(\Phi^{-1}(a))^{-1}$. In this example $\delta(x) = (\psi(\alpha_{20}(x)),-\sigma_{10}\psi(\alpha_{10}(x)))$. Therefore, from Theorem 3.1 we have $\zeta = \delta(x)\{y - E[y|x]\}$. Let $\omega(x) = \delta(x)\mathrm{Var}(y|x)\delta(x)' = \mathrm{Var}(\zeta|x)$. Then the optimal instruments are

$$\bar{a}(x) = D(x)'\omega(x)^{-1} = \begin{pmatrix} x_2-x_1 \\ \Phi^{-1}(\alpha_{10}(x)) \end{pmatrix}/\mathrm{Var}(\psi(\alpha_{20}(x))y_2-\sigma_{10}\psi(\alpha_{10}(x))y_1|x).$$

These instruments correspond to the first order condition for the weighted least squares estimator

$$\hat{\theta} = \mathrm{argmin}_\theta \sum_{i=1}^n \omega(x_i)^{-1}\rho(x_i,\theta,\hat{\alpha})^2/n$$

This estimator can also be thought of as a semiparametric minimum distance estimator, where θ is being chosen to minimize a function that should converge to zero at the truth. The characterization of an optimal IV estimator applies to any semiparametric minimum distance problem where $\alpha_0(x) = E[y|x]$ for a vector y and $\rho(x,\theta_0,\alpha_0) = 0$. For $D(x) = \partial\rho(x,\theta_0,\alpha_0)/\partial\theta$ and $\omega(x) = \delta(x)Var(y|x)\delta(x)'$, the optimal instruments will be $D(x)'\omega(x)^{-1}$. Furthermore, the weighted least squares estimator will be optimal in this class.

Construction of an efficient estimator is straightforward in the case where $x \subseteq w$. The optimal instruments in Theorem 2.1 can be estimated nonparametrically, proceeding analogously to Newey (1993). Alternatively, an approximately efficient estimator can be constructed by GMM estimation using a vector of approximating functions $A(x)$ as instruments, where the moment functions would be $A(x)\rho(z,\theta,\hat{\alpha})$. In this case, where the influence function is so simple, the spanning condition of Theorem 2.1 just requires that a linear combination of $A(x)$ can approximate the optimal instruments. For brevity we omit a formal result.

The next case we consider is that where $w \subseteq x$. This case is more complicated in that the correction for the first stage estimation does not lead to the adjusted residual form of the influence function. Let $\rho = \rho(z,\theta_0,\alpha_0)$, $\Omega(x) = E[\rho\rho'|x]$, $V = \delta(w)[d-\alpha_0(w)]$, $\Sigma(w) = E[VV'|w]$, and $K(x) = E[\rho V'|x]$.

Theorem 3.2: If $w \subseteq x$ *and Assumption 3.1 is satisfied, then*

$$u_m(z) = a(x)\rho + E[a(x)|w]V.$$

Also, if the linear equations

$$\bar{a}(x) = [D(x)' + P(w)' + R(w)'K(x)']\Omega(x)^{-1}.$$

$$P(w) = -\Sigma(w)E[\bar{a}(x)'|w] - E[K(x)'\bar{a}(x)'|w], \quad R(w) = -E[\bar{a}(x)'|w],$$

have a solution for $\bar{a}(x)$, $R(w)$, *and* $P(w)$ *with probability one, then the optimal instruments are* $\bar{a}(x)$. *Furthermore, if* $K(x) = 0$ *then* $R(w) = 0$ *and*

$$P(w) = -\{\Sigma(w)^{-1} + E[\Omega(x)^{-1}|w]\}^{-1}E[\Omega(x)^{-1}D(x)|w].$$

In general the form of the optimal instruments is quite complicated, although it simplifies in the zero conditional covariance case, $K(x) = 0$.

An example where Theorem 3.2 applies is a nonparametric generated regressor model where

$$\rho(z,\theta,\alpha) = y - z_1'\theta_1 - \gamma\alpha(w) - \delta[d - \alpha(w)], \quad \alpha_0(w) = E[d|w], \quad (d,w) \subseteq x.$$

This residual is that for a linear model, where a conditional expectation and the residual from the same conditional expectation are included as regressors. The model is a semiparametric version of a familiar model with many economic applications, that has been considered in Rilstone (1989). Note here that $D(x) = -(E[z_1'|x], \alpha_0(w), d-\alpha_0(w))$, $K(x) = 0$ and that Assumption 3.1 is satisfied for $\delta(w) = \delta_0 - \gamma_0$. Therefore, the optimal instruments are

$$\bar{a}(x) = D(x)'\Omega(x)^{-1} + E[D(x)'\Omega(x)^{-1}|w]\{\Sigma(w)^{-1} + E[\Omega(x)^{-1}|w]\}^{-1}\Omega(x)^{-1}.$$

In the case where $\Omega(x)$ and $\Sigma(w)$ are constant, this formula simplifies to

$$\bar{a}(x) = \begin{pmatrix} E[z_1 \mid x] \\ \alpha_0(w) \\ d-\alpha_0(w) \end{pmatrix} \Omega^{-1} + \frac{\Sigma}{\Sigma + \Omega} \begin{pmatrix} E[z_1 \mid w] \\ \alpha_0(w) \\ 0 \end{pmatrix} \Omega^{-1},$$

and it can be shown that the resulting estimator attains Rilstone's (1989) semiparametric bound for normal disturbances. It is interesting to note that these instruments are equal to those for the best estimator if $\alpha_0(w)$ were used in place of $\hat{\alpha}$, plus a term involving the additional variable $E[z_1|w]$. If $E[z_1|w] = E[z_1|x]$, e.g. as would occur if $z_1 \subseteq w$, then the best instruments would be a linear combination of the instruments that are best without the generated regressor problem. Indeed, if $z_1 \subseteq w$, then it follows that least squares is best.

In the constant $\Omega(x)$ and $\Sigma(w)$ case, an estimator of the optimal instruments is readily available from replacing $E[z_1|x]$, $E[z_1|w]$, $\alpha_0(w)$, Σ, and Ω by corresponding estimates. Alternatively, since the optimal instruments are a linear combination of $A(x) = (E[z_1'|x], E[z_1'|w], \alpha_0(w), d-\alpha_0(w))'$, for the corresponding $\hat{A}(x)$ an optimal estimator could be obtained from GMM with an optimal weighting matrix that accounts for the generated regressors.

In the general model of Theorem 3.2 it should be possible to construct an efficient estimator by using nonparametric estimates of the optimal instruments. Alternatively, an approximately efficient estimator can be constructed using many moment conditions. Here that corresponds to GMM estimation with many functions of x as instruments. It is straightforward to give conditions for approximate efficiency of these estimators. Let $A_J(x) = (A_{1J}(x),\ldots,A_{JJ}(x))'$ be a vector of functions of x, and consider a GMM estimator as in equation (10) or (11) with $m_J(z,\theta,\hat{\alpha}) = \rho(z,\theta,\hat{\alpha}) \otimes A_J(x)$. Let r be the dimension of ρ.

Theorem 3.3: If $E[u_m(z)u_m(z)']$ *is nonsingular,* $E[\|\bar{a}(x)\|^2] < \infty$, $\Omega(x)$ *and* $\Sigma(w)$ *are bounded, and for any* $a(x)$ *with* $E[\|a(x)\|^2] < \infty$, *there exists* C_J *such that* $E[\|a(x)-C_J\{I_r \otimes A_J(x)\}\|^2] \longrightarrow 0$ *as* $J \longrightarrow \infty$, *then* $(H_J'V_J^{-1}H_J)^{-1} \longrightarrow (E[u_m(z)u_m(z)'])^{-1}$ *as* $J \longrightarrow \infty$.

The point of this result is that all that is needed to guarantee approximation of the complicated optimal influence function in Theorem 3.2, leading to approximate efficiency, is that the instruments can approximate functions of x. There are many types of instruments that would meet this qualification, including power series and regression splines, making it relatively easy to construct an approximately efficient estimator.

Here we have shown the form of an efficient two-step GMM estimator when the second step instruments are a subset of the first step regressors or vice versa. It is also possible to obtain some results for the case where w and x are not a subset of each other. For brevity these results are omitted.

4. Sample Selection Models and Nonparametric Propensity Score Estimation

Interesting and important two-step estimators arise in the context of sample selection models. A general form of such a model is

$$y^* = x'\theta_0 + \varepsilon, \quad y^* \text{ only observed if } d = 1, \quad d \in \{0, 1\}. \tag{16}$$

$$\text{Prob}(d = 1|w) = \pi(w) = P, \quad x \subseteq w.$$

The selection probability P is referred to as the propensity score. The parameters of interest θ_0 are identified under various restrictions on the joint distribution of the selection indicator d and the disturbance ε. In this section we consider the form of

optimal two-step estimators in two cases, where ε and w are mean independent conditional on $d = 1$ and P and where they are statistically independent. The estimators are two-step estimators where the first step is nonparametric estimation of the propensity score. Some such estimators have been previously considered by Ahn and Powell (1993) and Choi (1990).

The first case we consider is

$$E[\varepsilon|w,d=1] = E[\varepsilon|P,d=1], \tag{17}$$

Here the conditional mean of the disturbance, given selection and w, depends only on the propensity score. This can be motivated by a latent variable model where $d = 1(\tau(w)+\eta \geq 0)$ for some unknown function $\tau(w)$ and a disturbance η. Suppose that P is a one-to-one function of $\tau(w)$ (e.g. η and w are independent and η has density that is positive everywhere) and that $E[\varepsilon|\eta,w] = E[\varepsilon|\eta,\tau(w)]$. Then $E[\varepsilon|w,d=1] = E[\varepsilon|w,\eta\geq-\tau(w)] = E[E[\varepsilon|\eta,w]|w,\eta\geq-\tau(w)] = E[E[\varepsilon|\eta,\tau(w)]|w,\eta\geq-\tau(w)]$, which is a function of w only through $\tau(w)$, and hence through P, so that equation (17) holds.

A two step estimator of θ_0 can be constructed using a nonparametric propensity score estimator $\hat{P} = \hat{\pi}(w)$. A vector of instrumental variables $a(w)$ can be used along with a residual $d\{y-\hat{E}[y|\hat{P},d=1] - (x-\hat{E}[x|\hat{P},d=1])'\theta\}$, where $\hat{E}[\cdot|\hat{P}]$ is a nonparametric regression estimator with regressor $\hat{P}$, to form a moment vector

$$m(z,\theta,\alpha) = a(w)d\{y-\alpha_y(\alpha_\pi(w)) - [w-\alpha_w(\alpha_\pi(w))]'\theta\}, \quad \alpha = (\alpha_y, \alpha_w, \gamma). \tag{18}$$

Here α_y and α_w have true values $E[y|\alpha_\pi(w)]$ and $E[w|\alpha_\pi(w)]$, and $\alpha_\pi(w)$ has true value $\pi(w)$. Using this function with $\hat{\alpha}$ as we have described is a two-step instrumental variables version of Robinson's (1988) estimator, and a density weighted version has been developed by Ahn and Powell (1993).

It is straightforward to derive the influence function from the results of Newey

(1994) and to obtain the optimal instruments. Let $\lambda(P) = E[\varepsilon|w,d=1]$ and $\lambda_P = \partial\lambda(P)/\partial P$.

Theorem 4.1: The influence function is

$$u_m(z) = \{a(w)-E[a(w)|P]\}\zeta, \quad \zeta = d[\varepsilon-\lambda(P)] + P\lambda_P(P)(d-P).$$

For $\eta(w)^2 = E[\zeta^2|w] = E[d\{\varepsilon-\lambda(P)\}^2|w] + \lambda_P(P)^2P^3(1-P)$ *bounded away from zero, the optimal instruments are*

$$\bar{a}(w) = P\eta(w)^{-2}\{x - E[\eta(w)^{-2}x|P]/E[\eta(w)^{-2}|P]\}.$$

Furthermore, $Var(u_m(z))^{-1}$ *is the semiparametric variance bound for estimation of* θ_0 *in the model of equation (17).*

The best instruments here are like those that appear in an efficient estimator for a heteroskedastic partially linear model as discussed in Chamberlain (1992). They are obtained from partialling out an unknown function of P in a weighted least squares criterion, where the weight is the inverse of the conditional variance of ζ. The main difference with Chamberlain (1992) is that the presence of $\hat{P}$ leads to the weight being the inverse conditional variance of the adjusted residual ζ, rather than the original residual ρ. In this way the optimal instruments account for the presence of the first stage nonparametric estimates.

Because the optimal instruments attain the semiparametric efficiency bound we do not need to search beyond instrumental variables estimation for an efficient estimator. Construction of an optimal estimator, or approximately optimal estimator could be carried similar to the way discussed in Section 2. A nonparametric estimator $\hat{\eta}(w)^2 = \hat{E}[d\{y-x'\hat{\theta}-\hat{\lambda}(\hat{P})\}^2|w] + \hat{\lambda}_P(\hat{P})^2\hat{P}^3(1-\hat{P})$ could be constructed using $\hat{\theta}$ and $\hat{\lambda}$ from an initial unweighted least squares estimator of a partially linear model $y = x'\theta + \lambda(\hat{P}) + r$ in the selected data, and then a nonparametric estimator of the optimal instruments

formed as

$$\hat{a}(w) = \hat{P}\hat{\eta}(w)^{-2}\{x - \hat{E}[\hat{\eta}(w)^{-2}x|\hat{P}]/\hat{E}[\hat{\eta}(w)^{-2}|\hat{P}]\}.$$

This is a complicated estimator for which regularity conditions have not yet been formulated in the literature, but should lead to efficiency.

An approximately efficient estimator is straightforward to construct here, by using approximating functions as instruments and then doing optimal GMM that accounts for the presence of the nonparametric first stage. Let $A_J(w)$ be a vector of approximating functions and consider a GMM estimator as in equation (10) or (11) with $m_J(z,\theta,\hat{\alpha}) = A_J(w)d\{y-\hat{E}[y|\hat{P},d=1] - (x-\hat{E}[x|\hat{P},d=1])'\theta\}$. Then there is a relatively simple spanning condition for approximate efficiency of the GMM estimator:

Theorem 4.2: If $E[u_m(z)u_m(z)']$ *is nonsingular,* $\eta(w)$ *is bounded and for any* $a(w)$ *with finite mean-square, there exists* C_J *such that* $E[\|a(w)-C_JA_J(w)\|^2] \longrightarrow 0$ as $J \longrightarrow \infty$, *then* $(H_J'V_J^{-1}H_J)^{-1} \longrightarrow (E[u_m(z)u_m(z)'])^{-1}$ *as* $J \longrightarrow \infty$.

Here we see that it suffices for approximate efficiency that the approximating vector $A_J(w)$ spans the set of functions with finite mean-square. There are many such functions that could be used, including splines and power series.

The second sample selection model we consider satisfies equation (16) with

$$\varepsilon \text{ and } w \text{ are independent conditional on } P \text{ and } d = 1. \tag{19}$$

Here it is assumed that conditioning on P removes any dependence between ε and w. This can be motivated by a latent variable model like that above, where $d = 1(\tau(w)+\eta \geq 0)$. If the joint distribution of (ε,η) given w depends only on $\tau(w)$ then this equation will be satisfied.

A basic implication of the conditional independence in equation (19) is that any function of w will be uncorrelated with any function of ε and P, conditional on P

and $d = 1$. This allows us to form estimators analogous to those above, where $y-x'\theta$ is replaced by any function of $y - x'\theta$ and $\hat{P}$. To be precise, for a vector of instrumental variables $a(w)$, a function $q(\varepsilon,P)$, and a corresponding residual $d\{q(y-x'\theta,\hat{P}) - \hat{E}[q(y-x'\theta,\hat{P})|\hat{P},d=1]\}$, we can use the moment function

$$m(z,\theta,\alpha) = a(w)d[q(y-x'\theta,\alpha_\pi(w)) - \alpha_q(y-x'\theta,\alpha_\pi(w))], \quad \alpha = (\alpha_q, \alpha_\pi). \qquad (20)$$

Here α_q and α_π have true values $E[q(y-x'\theta,\alpha_\pi(w))|\alpha_\pi(w),d=1]$ and $\pi(w)$, respectively. Using this function with $\hat{\alpha}$ as we have described is a nonlinear version of the estimator described above for the conditional mean case.

The next result gives the form of the influence function and the optimal moment functions for this estimator. Here let $\lambda(P) = E[q(\varepsilon,P)|P,d=1]$, $\lambda_P(P) = \partial\lambda(P)/\partial P$, and $f(\varepsilon|P)$ the density of ε given $d = 1$ and w. Also, set $s_\varepsilon(\varepsilon,P) = d\ln f(\varepsilon|P)/d\varepsilon$ and $s_P(\varepsilon,P) = d\ln f(\varepsilon|P)/dP$ be the score for f with respect to a location parameter and P.

Theorem 4.3: For the model of equation (19) and moment functions as in equation (20), the influence function is

$$u_m(z) = \{a(w)-E[a(w)|P,d=1]\}\zeta, \quad \zeta = d[q(\varepsilon,P)-\lambda(P)] + P\{E[q_P(\varepsilon,P)|P,d=1]-\lambda_P(P)\}(d-P),$$

and the optimal choice of moment function has $\bar{a}(w) = x$ *and*

$$\bar{q}(\varepsilon,P) = s_\varepsilon(\varepsilon,P) - Ps_P(\varepsilon,P)E[s_\varepsilon s_p|P,d=1]/\{P^{-1}(1-P)^{-1} + PE[s_p^2|P,d=1]\}.$$

Furthermore, $Var(u_{\bar{m}}(z))^{-1}$ *is the semiparametric variance bound for estimation of* θ_0.

Using many moment functions in an approximately efficient estimator is useful in this case, where the best $q(\varepsilon,P)$ is quite complicated, but the optimal instrument $\bar{a}(w) = x$ has a known functional form. In this case approximate efficiency can be achieved with only a two-dimensional approximation of the best $q(\varepsilon,P)$, rather than the potentially high dimensional approximation of the best instruments $\bar{a}(w)$ from Theorem

4.1. The low dimensional nature of the approximation means that it should be possible to attain high efficiency using a relatively small set of moment conditions. Let $q_J(\varepsilon,P)$ denote a vector of approximating functions, and consider a GMM estimator as in equation (10) or (11) with $m_J(z,\theta,\hat{\alpha}) = d\{q_J(y-x'\theta,\hat{P}) - \hat{E}[q_J(y-x'\theta,\hat{P})|\hat{P},d=1]\}\otimes x$.

Theorem 4.4: If $E[u_m(z)u_m(z)']$ *is nonsingular,* x *and* $E[(s^P)^2|P,d=1]$ *are bounded, and for any* $E[d\tilde{q}(\varepsilon,P)^2]$ *finite there is* C_J *such that* $E[d\{\tilde{q}(\varepsilon,P)-C_Jq_J(\varepsilon,P)\}^2] \longrightarrow 0$ *as* $J \longrightarrow \infty$, *then* $(H_J'V_J^{-1}H_J)^{-1} \longrightarrow (E[u_m(z)u_m(z)'])^{-1}$ *as* $J \longrightarrow \infty$.

In practice it may make sense to begin the approximation with functions of ε and P that correspond to particular distributions. For example, one could derive the form of $\bar{q}(\varepsilon,P)$ when ε and η have a normal distribution. Alternatively, one could use a few simple approximating functions, like power series in some function of ε. Because selection is likely to induce some skewness and because normality is not expected in many econometric applications, using such an estimator for the conditional independence case of equation (19) can result in substantial efficiency gains over the linear estimators based on equation (18), as shown in Newey (1991) for a semiparametric selection probability.

APPENDIX: Proofs of Theorems

Let C denote a constant that is different in different uses.

Proof of Theorem 2.1: For a matrix B let $\|B\| = [\mathrm{tr}(B'B)]^{1/2}$. By equation (9), $H_J = E[u_J \underline{u}_m']$, where we suppress the z argument for notational convenience, so that $\tilde{\theta}_J$ has asymptotic variance

$$(H_J' V_J^{-1} H_J)^{-1} = (E[\underline{u}_m u_J'](E[u_J u_J'])^{-1} E[u_J \underline{u}_m'])^{-1} = (E[\tilde{u}_J \tilde{u}_J'])^{-1},$$

$$\tilde{u} = E[\underline{u}_m u_J'](E[u_J u_J'])^{-1} u_J.$$

For the C_J in the statement of the result, let $u_C = C_J u_J$. Since $\tilde{u}$ is the multivariate least squares projection of $\underline{u}_m$ on u_J, it follows that

$$E[u_C u_C'] \leq E[\tilde{u}\tilde{u}'] \leq E[\underline{u}_m \underline{u}_m']. \tag{21}$$

Also, by the spanning condition in the statement,

$$\|E[u_C u_C'] - E[\underline{u}_m \underline{u}_m']\| \leq E[\|u_C u_C' - \underline{u}_m \underline{u}_m'\|]$$

$$\leq E[\|u_C - \underline{u}_m\|^2] + 2(E[\|u_C - \underline{u}_m\|^2])^{1/2}(E[\|\underline{u}_m\|^2])^{1/2} \longrightarrow 0.$$

Therefore, by equation (21), $E[\tilde{u}\tilde{u}'] \longrightarrow E[\underline{u}_m \underline{u}_m']$ as $J \longrightarrow \infty$, so the conclusion follows by nonsingularity of $E[\underline{u}_m \underline{u}_m']$ and continuity of the inverse matrix at any nonsingular matrix. QED.

Proof of Theorem 3.1: The form of the influence function follows by Newey (1994). The optimal instruments follow as in Chamberlain (1987).

Proof of Theorem 3.2: The form of the influence function follows by Newey (1994). For notational convenience suppress the x argument of $a(x)$, $\Omega(x)$, $D(x)$, and $K(x)$.

Then by iterated expectations, equation (9) for the optimal influence function, and hence the optimal instruments, is

$$H_m = E[aD] = E[a\{\Omega\bar{a}' + \Sigma(w)E[\bar{a}'|w] + KE[\bar{a}'|w] + E[K'\bar{a}'|w]\}].$$

Since $a = a(x)$ can be any bounded function of x, satisfaction of this equation requires that $D = \Omega\bar{a}' + \Sigma(w)E[\bar{a}'|w] + KE[\bar{a}'|w] + E[K'\bar{a}'|w] = \Omega\bar{a}' - P(w) - KR(w)$. Solving for $\bar{a}$ then gives the the second result. For the third result, let $P_w = P(w)$ and $\Sigma_w = \Sigma(w)$. Applying the optimal instrument formula gives $\bar{a} = (D' + P_w')\Omega^{-1}$, $P_w = -\Sigma_w E[\bar{a}'|w]$. Plugging in the formula for $\bar{a}$ in the equation for P_w gives $P_w = -\Sigma_w E[\Omega^{-1}(D + P_w)|w] = -\Sigma_w E[\Omega^{-1}D|w] - \Sigma_w E[\Omega^{-1}|w]P_w$. Solving for P_w then gives the third result. QED.

Proof of Theorem 3.3: Let $\tilde{a}_J(x) = I_r \otimes a_J(x)$. Note that for any $u_m(z)$ as in Theorem 3.1 or Theorem 3.2, and conformable constant matrix C_J,

$$E[\|u_m(z) - u_{C_J m_J}(z)\|^2]$$

$$\leq 2E[\|a(x) - C_J\tilde{a}_J(x)\}\|^2\|\Omega(x)\|] + 2E[\|E[\{a(x) - C_J\tilde{a}_J(x)\}|w]\|^2\|\Sigma(w)\|]$$

$$\leq CE[\|a(x) - C_J\tilde{a}_J(x)\}\|^2] + CE[E[\|a(x) - C_J\tilde{a}_J(x)\|^2|w]] \leq CE[\|a(x) - C_J\tilde{a}_J(x)\}\|^2].$$

The conclusion then follows by Theorem 2.1. QED.

For the proof of Theorems 4.1 - 4.4, for any function b of the data let $E_P b = E[b|P,d=1]$, and note that for functions $b(w)$ of w, $E_p b = E[b|P]$.

Proof of Theorem 4.1: As shown in Newey (1994), the correction terms for nonparametric estimation can be derived separately for α_y, α_w, and α_π. From the form of the conditional expectations correction in Newey (1994) it follows that the correction term for α_y and α_w is $-E[a(w)|P,d=1]\rho$. Also, let $\pi(w,\gamma) = E_\gamma[d|w]$ denote the

propensity score when for a distribution parameterized by γ that passes through the truth. Then by Newey (1994) the correction term for nonparametric estimation of $\pi(w)$ can be computed from the derivative of $E[da(w)E[\varepsilon|\pi(w,\gamma),d=1]]$ with respect to γ at the truth. By equation (21), iterated expectations, the chain rule, and the fact that $\partial\pi(w,\gamma_0)/\partial\gamma = \partial E_\gamma[d|w]/\partial\gamma = E[\{d-P\}S_\gamma(z)|w]$ for the score $S_\gamma(z)$ for γ,

$$\partial E[da(w)E[\varepsilon|\pi(w,\gamma),d=1]]/\partial\gamma = \partial E[da(w)E[\lambda(P)|\pi(w,\gamma),d=1]]/\partial\gamma$$

$$= \partial E[da(w)E[\lambda(P-\pi(w,\gamma)+P)|P,d=1]]/\partial\gamma + \partial E[da(w)\lambda(\pi(w,\gamma))]/\partial\gamma$$

$$= E[P\lambda_P(P)\{a(w)-E[a(w)|P]\}\partial\pi(w,\gamma)/\partial\gamma] = E[P\lambda_P(P)\{a(w)-E[a(w)|P]\}\{d-P\}S_\gamma(z)].$$

Then by Newey (1994) the correction term for estimation of P is $P\lambda_P(P)\{a(w)-E[a(w)|P]\}(d-P)$. Noting that $E[a(w)|P,d=1] = E[da(w)|P]/E[d|P] = E[a(w)|P]$, we obtain the first conclusion. Also, $H_m = E[\partial m(z,\theta,\alpha_0)/\partial\theta] = -E[Pa(w)\{x-E[x|P]\}']$. Equation (9) is then

$$-E[aP(x-E_Px)'] = -E[(a-E_Pa)Px'] = E[(a-E_Pa)\zeta^2(\bar{a}-E_P\bar{a})'] = E[(a-E_Pa)\eta^2(\bar{a}-E_P\bar{a})'],$$

where the w argument is suppressed for convenience. Subtracting gives

$$-E[(a-E_Pa)(Px-\eta^2\{\bar{a}-E_P\bar{a}\})'] = 0.$$

Since $a(w)$ can be any function of w, this equation implies that

$$Px-\eta^2(\bar{a}-E_P\bar{a}) = h(P)$$

for some function $h(P)$ of P. Dividing through by η^2 and taking conditional expectations given P gives $PE_P\eta^{-2}x = h(P)E_P\eta^{-2}$. Solving, $h(P) = PE_P\eta^{-2}x/E_P\eta^{-2}$, and solving again,

$$\bar{a} - E_P\bar{a} = P\eta^{-2}[x - E_P(\eta^{-2}x)/E_P(\eta^{-2})] = P\eta^{-2}(x - E[\eta^{-2}x|P]/E[\eta^{-2}|P]).$$

Setting $\bar{a}$ equal to the expression on the right-hand side, and noting that $E[\bar{a}|P] = 0$ for that choice, gives the second result. The last result follows from equation (3.20) of Newey and Powell (1993). QED.

Proof of Theorem 4.2: Choose $A_J = A_J(w)$ and C_J such that for $a_J = C_J A_J$, $E[\|\bar{a}-a_J\|^2] \longrightarrow 0$. Then since $E_P\bar{a} = 0$, $E[\|\underline{u}_m(z)-C_J u_J(z)\|^2] = E[\|\{(\bar{a}-a_J)-E_P(\bar{a}-a_J)\}\zeta\|^2] \leq CE[\|\bar{a}-a_J\|^2\eta^2] + CE[\|E_P(\bar{a}-a_J)\|^2\eta^2] \leq CE[\|\bar{a}-a_J\|^2] + CE[E_P\|\bar{a}-a_J\|^2] \leq CE[\|\bar{a}-a_J\|^2] \longrightarrow 0$. The conclusion then follows by Theorem 2.1. QED.

Proof of Theorem 4.3: It follows as for the conditional mean case that the correction term for estimation of α_q is $-E[a(w)|P,d=1]d\{q(\varepsilon,P) - \lambda(P)\}$, where $\lambda(P) = E[dq(\varepsilon,P)|P,d=1]$. Similarly, the correction term for estimation of P is $\{a(w)-E[a(w)|P,d=1]\}P\{E[q_P(\varepsilon,P)|P]-\lambda_P(P)\}(d-P)$, where P and ε subscripts will denote corresponding partial derivatives, giving the first conclusion.

To solve equation (9) for the optimal moment functions, note $H_m = E[\partial m(z,\theta,\alpha_0)/\partial\theta]$ $= -E[a(w)\{dq_\varepsilon(\varepsilon,P)x-E[q_\varepsilon(\varepsilon,P)x|P,d=1]\}]$. Let $f(\varepsilon|P)$ denote the density of ε given w and $d = 1$, $s_\varepsilon = s_\varepsilon(\varepsilon,P)$, and $s_P = s_P(\varepsilon,P)$. For notational simplicity let $q_\varepsilon = q_\varepsilon(\varepsilon,P)$ and $q_P = q_P(\varepsilon,P)$. Note that $E[dq_\varepsilon|w] = PE[dq_\varepsilon|w,d=1] + (1-P)E[dq_\varepsilon|w,d=0] = PE[dq_\varepsilon|w,d=1] = PE_Pq_\varepsilon$. Therefore,

$$H_m = -E[aP(E_Pq_\varepsilon)(x-E_Px)'] = -E[(a-E_Pa)P(E_Pq_\varepsilon)(x-E_Px)'].$$

Let $\rho = d[q(\varepsilon,P)-\lambda(P)]$. Integration by parts of $E_Pq_\varepsilon = \int q_\varepsilon(\varepsilon,P)f(\varepsilon|P)d\varepsilon$, differentiating $\lambda(P) = \partial\int q(\varepsilon,P)f(\varepsilon|P)d\varepsilon/\partial P$ with respect to P, and using $E_Ps_\varepsilon = E_Ps_P = 0$, we obtain

$$E_Pq_\varepsilon = \int q_\varepsilon(\varepsilon,P)f(\varepsilon|P)d\varepsilon = -E_Pq(\varepsilon,P)s_\varepsilon = -E_P(\rho s_\varepsilon), \quad E_Pq_P - \lambda_P(P) = -E_P(\rho s_P).$$

It follows that

$$\zeta = \rho - E_P(\rho s_P)P(d-P).$$

Let $\bar{a}$ and $\bar{q}$ denote the optimal functions, $\bar{\rho} = d\{\bar{q}(\varepsilon,P)-E_P\bar{q}(\varepsilon,P)\}$, and $\bar{\zeta} = \bar{\rho} - E_P(\bar{\rho}s_P)P(d-P)$. Note that by conditional independence, $E[\rho\bar{\rho}\,|\,w] = PE_P(\rho\bar{\rho})$. Then equation (9) is

$$E[(a-E_Pa)PE_P(\rho s_\varepsilon)(x-E_Px)'] = H_m = E[u_m(z)\underline{u}_m(z)']$$

$$= E[(a-E_Pa)E[\zeta\bar{\zeta}\,|\,w](\bar{a}-E_P\bar{a})'] = E[P(a-E_Pa)\{E_P(\rho\bar{\rho}) + P^2(1-P)E_P(\rho s_P)E_P(\bar{\rho}s_p)\}(\bar{a}-E_P\bar{a})'].$$

As in the proof of Theorem 4.1, equality for all $a(w)$ requires that

$$E_P(\rho s_\varepsilon)(x-E_Px)' - \{E_P(\rho\bar{\rho}) + P^2(1-P)E_P(\rho s_P)E_P(\bar{\rho}s_p)\}(\bar{a}-E_P\bar{a})' = h(P).$$

Taking conditional expectations of both sides given P we find that $h(P) = 0$. This equation will hold if $\bar{a} = x$, and for $g(P) = E_P(\bar{\rho}s_p)$,

$$0 = E_P(\rho s_\varepsilon) - [E_P(\rho\bar{\rho}) + P^2(1-P)E_P(\rho s_P)g(P)]$$

$$= -E[dq(\varepsilon,P)\{\bar{\rho} - [s_\varepsilon - P^2(1-P)s_Pg(P)]\}\,|\,P,d=1].$$

For this equality to hold for any function $q(\varepsilon,P)$ it is sufficient that $\bar{q} = [s_\varepsilon - P^2(1-P)s_Pg(P)]$, since $E_P\bar{q} = 0$. Multiplying through by s_P and taking a conditional expectation gives $g(P) = E_P(s_\varepsilon s_P) - P^2(1-P)E_Ps_P^2g(P)$. Solving, $g(P) = E_P(s_\varepsilon s_P)/[1 + P^2(1-P)E_Ps_P^2]$. An optimal q is then given by

$$\bar{q}(\varepsilon,P) = s_\varepsilon - g(P)s_P, \quad g(P) = PE_p(s_\varepsilon s_p)/[P^{-1}(1-P)^{-1} + PE_p(s_p^2)].$$

To show the last conclusion, note that $E_P\bar{q} = 0$, so that $\bar{\rho} = d\bar{q}$. Also, $E_P(\bar{\rho}s_P) = g(P)P^{-2}(1-P)^{-1}$, so that

$$\bar{\zeta} = d\bar{q} - E_P(\bar{\rho}s_P)P(d-P) = d\bar{q} - g(P)P^{-1}(1-P)^{-1}(d-P).$$

Therefore, $u_{\underline{m}}(z) = (x-E_P x)\bar{\zeta}$ matches the efficient score given in equation (4.15) of Newey and Powell (1993), giving the last conclusion. QED.

Proof of Theorem 4.4: By arguments like those above, $u_J(z) = (x - E[x|P])\{\rho_J - E[\rho_J s^P|P](d - P)\}$ for $\rho_J = d(p_J - E[p_J|P,d=1])$. Suppose that there is $\tilde{C}_J$ such that as $J \longrightarrow \infty$, $E[dw(x)\{\bar{p}(\varepsilon,P)-\tilde{C}_J' p_J(\varepsilon,P)\}^2] = o(1)$ for $w(x) = \|x-E[x|P]\|^2 + P(1-P)E[(s^P)^2|P]E[\|x-E[x|P]\|^2|P]$. Then for $C_J = \tilde{C}_J \otimes I_q$ and $\epsilon_J = \bar{p}(\varepsilon,P)-\tilde{C}_J' p_J(\varepsilon,P)$, where q is the dimension of x, $E[\|u_{\underline{m}}(z)-C_J' u_J(z)\|^2] = E[\|x-E[x|P]\|^2\{\bar{\rho}-\tilde{C}_J'\rho_J\}^2] + 2E[P(1-P)\|x-E[x|P]\|^2\{E[\{\rho-\tilde{C}_J'\rho_J(\varepsilon,P)\}s^P|P]\}^2] \leq E[dw(x)\{\epsilon_J - E[\epsilon_J|P,d=1]\}^2] \leq 2E[dw(x)\epsilon_J^2] + 2E[dw(x)\{E[\epsilon_J|x,d=1]\}^2] \leq$

$4E[dw(x)\epsilon_J^2] = o(1)$.

References

Ahn, H. and C.F. Manski, 1993, Distribution theory for the analysis of binary choice under uncertainty with nonparametric estimation of expectations, Journal of Econometrics 58, 291-321.

Ahn, H. and J.L. Powell, 1993, Semiparametric estimation of censored selection Models with a nonparametric selection mechanism, Journal of Econometrics 58, 3-29.

Bates, C.E. and H. White, 1993, Determination of estimators with minimum asymptotic covariance matrices, Econometric Theory 9, 633-648.

Beran, R., 1976, Adaptive estimates for autoregressive processes, Annals of the Institute of Statistical Mathematics 26, 77-89.

Chamberlain, G., 1987, Asymptotic efficiency in estimation with conditional moment restrictions, Journal of Econometrics 34, 305-334.

Chamberlain, G., 1992, Efficiency bounds for semiparametric regression, Econometrica 60, 567-596.

Choi, K., 1990, The semiparametric estimation of the sample selection model using series expansion and the propensity score, manuscript, Department of Economics, University of Chicago.

Crepon, B., F. Kramarz, and A. Trognon, 1997, Parameters of interest, nuisance parameters, and orthogonality conditions: An application to autoregressive error component models, Journal of Econometrics 82, 135-156.

Hansen, L. P., 1982, Large sample properties of generalized method of moments estimators, Econometrica 50, 1029-1054.

Hansen, L.P., 1985a, A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators, Journal of Econometrics 30, 203-238.

Hansen, L.P., 1985b, Two-step generalized method of moments estimators, discussion, North American Winter Meeting of the Econometric Society, Meeting, New York.

Hansen, L.P., J.C. Heaton, and M. Ogaki, 1988, Efficiency bounds implied by multiperiod conditional moment restrictions, Journal of the American Statistical Association 83, 863-871.

Hayashi, F., and C. Sims, 1983, Nearly efficient estimation of time series models with predetermined, but not exogenous, instruments, Econometrica 51, 783-798.

Newey, W.K., 1990, Semiparametric efficiency bounds, Journal of Applied Econometrics 5, 99-135.

Newey, W.K., 1991, Two-step series estimation of sample selection models, working paper, MIT Department of Economics.

Newey, W.K., 1993, Efficient estimation of models with conditional moment restrictions, in G.S. Maddala, C.R. Rao, and H.D. Vinod, eds., Handbook of Statistics, Volume 11: Econometrics. Amsterdam: North-Holland.

Newey, W.K., 1994, The asymptotic variance of semiparametric estimators, Econometrica 62, 1349-1382.

Newey, W.K. and D. McFadden, 1994, Large Sample Estimation and Hypothesis Testing, R. Engle and D. McFadden (eds.), Handbook of Econometrics, Vol. 4, Amsterdam, North-Holland, 2113-2245.

Newey, W.K. and J.L. Powell, 1993, Efficiency bounds for semiparametric selection models, Journal of Econometrics 58, 169-184.

Rilstone, P., 1989, Computing the (local) efficiency bound for a semiparametric generated regressors model, manuscript, Department of Economics, University of Western Ontario.

Robinson, P.M., 1988, Root-n consistent semiparametric regression, Econometrica 56, 931-954.

7078 006

Date Due

Lib-26-67

MIT LIBRARIES
3 9080 01972 0900